Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3153–3161
Marseille, 11–16 May 2020

Constructing Web-Accessible Semantic Role Labels and Frames for Japanese as Additions to the NPCMJ Parsed Corpus

Koichi Takeuchi*, **Alastair Butler**⋆, **Iku Nagasaki**†, **Takuya Okamura***, **Prashant Pardeshi**‡
*Okayama University, ⋆Hirosaki University, †Nagoya University,
‡National Institute of Japanese Language and Linguistics
{takeuc-k@, t-okamura@s}.okayama-u.ac.jp, ajb129@hirosaki.ac.jp,
inagasaki@free.fr, prashant@ninjal.ac.jp

Abstract

As part of constructing the NINJAL Parsed Corpus of Modern Japanese (NPCMJ), a web-accessible language resource, we are adding frame information for predicates, together with two types of semantic role labels that mark the contributions of arguments. One role type consists of numbered semantic roles, like in PropBank, to capture relations between arguments in different syntactic patterns. The other role type consists of semantic roles with conventional names. Both role types are compatible with hierarchical frames that belong to related predicates. Adding semantic role and frame information to the NPCMJ will support a web environment where language learners and linguists can search examples of Japanese for syntactic and semantic features. The annotation will also provide a language resource for NLP researchers making semantic parsing models (e.g., for AMR parsing) following machine learning approaches. In this paper, we describe how the two types of semantic role labels are defined under the frame based approach, i.e., both types can be consistently applied when linked to corresponding frames. Then we show special cases of syntactic patterns and the current status of the annotation work.

Keywords: semantic roles, predicate frames, sentence level meaning, PropBank, thesaurus

1. Introduction

Semantic role labeled data is important to capture predicate-argument sentence level meaning for NLP researchers, linguists, and language learners. For English, various kinds of annotated corpora with semantic roles and frames are provided (e.g., PropBank (Bonial et al., 2010) and FrameNet (Baker et al., 1998)) and they are widely used. The numbered semantic roles (e.g., Arg0, Arg1, Arg2) proposed in PropBank are effective for adapting to various systems of semantic roles. Thus, the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) approach, which is based on PropBank numbered semantic roles and frames, has attracted a lot of attention from NLP researchers as a model for describing sentence-level semantics. AMR parsing models are studied with deep neural network models, e.g., (Zhang et al., 2019) using AMR resources for English[1].

In terms of Japanese language resources, several corpora containing annotated tags related to predicate-argument information have been proposed (Kyoto (Kawahara et al., 2002), NAIST (Iida et al., 2007), EDR (EDR, 1995), GDA (Hashida, 2005), Japanese FrameNet (Ohara et al., 2006), BCCWJ-PT (Takeuchi et al., 2015)). However, numbered semantic roles have not been annotated for Japanese texts. Also, none of the annotated corpora that are reported in the literature are web accessible because of licensing restrictions[2].

In this research project, we annotate two types of semantic role labels and frames of predicates from the NINJAL Parsed Corpus of Modern Japanese (NPCMJ), which is a freely available web-accessible treebank (NINJAL, 2016)[3]. The semantic roles and frames are annotated based on a freely available web-accessible thesaurus of predicate-argument structure for Japanese (PT)[4]. Utilizing the tree structure of the NPCMJ, we can focus directly on the problem of labeling for the semantic roles and frames of target predicates. This is because the identification of target predicates and their arguments can be automatically derived from the treebank following (Horn et al., 2018).

The first type of annotation for semantic roles is PropBank-style, that is, numbered semantic roles. The second type involves names for semantic roles (e.g., Agent, Theme, Goal), which has been used in PT. The reason why we employ the conventional named semantic roles is that the named semantic roles are intuitively more understandable for humans compared to the numbered roles. Thus, the named roles are expected to be used in queries when language learners or linguists want to extract example sentences that, e.g., contain "Experiencer" arguments in the emotional sense, or "Source" arguments in the moving sense.

Contributions of this paper are as follows:

(1) For NLP researchers, linguists, and Japanese language learners, we propose a dual semantic role annotation that is composed of: (i) numbered semantic roles, and (ii) semantic roles with conventional names.

[1] LDC2017T10 and LDC2014T12.

[2] Japanese FrameNet is constructed on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014), from which it inherits release issues. Thus, the annotated sentences are not available on the website: http://sato.fm.senshu-u.ac.jp/frameSQL/jfn23/notes/index2.html (accessed 2019/11/22).

[3] The treebank annotation manual, the searching tool and the data are available from http://npcmj.ninjal.ac.jp/ (accessed 2019/11/30).

[4] PT: Predicate-argument Thesaurus. http://pth.cl.cs.okayama-u.ac.jp/testp/pth/Vths.

(2) The proposed numbered semantic roles are frame-based so that roles can be assigned consistently with regards to their corresponding frame.

(3) We reveal that the frame-based semantic roles are suitable for Japanese because of dividing transitive and intransitive verbs.

(4) We show that the current annotated predicates are about 10,000, and reveal highly promising accuracy results from a semantic role annotation system with a neural network model.

2. Frame-Based Semantic Roles

2.1. Frame-based Semantic Roles Are Needed in an Agglutinative Language Such as Japanese

With semantic role annotation, the aim is to fix the semantic relations between arguments and predicates in a manner that is abstracted away from differences between case markers and syntactic alternations. One of the difficulties of annotating semantic roles with names involves settling on a system of names, such as Agent, Theme, and Patient. Several different systems of semantic roles are proposed in previous studies (Fillmore, 1968; Palmer et al., 2010; Loper et al., 2007; Baker et al., 1998). By contrast, the numbered semantic roles proposed in PropBank can absorb these differences, because numbered arguments can be consistently assigned in example sentences for the target verb sense.

In Japanese, some verbs, e.g., 'hirak-u' (open), behave as both intransitive and transitive verbs without any intransitivizing or transitivizing suffix[5].

(5)	a.	doa ga	hirak-u	
		door NOM	open-PRS	
		Arg1		
		'The door opens'		
	b.	ken ga	doa o	hirak-u
		Ken NOM	door ACC	open-PRS
		Arg0	Arg1	
		'Ken opens the door'		

The 'doa' (door) takes a different case marker in intransitive use and transitive use, but both 'doa' are the opened object. Numbered semantic roles can be adapted for any system of named semantic roles. For example, the semantic role of 'doa' can be Theme or Patient, but numbered roles fix this relation by giving the unique number of the argument in the example sentence. These role combinations (called rolesets in PropBank) should be different according to the meaning of the predicates. Thus, in PropPank, each roleset is defined on each meaning of a predicate.

A notable difference of Japanese from English, that can be an obstacle for second language-learners, is that most verbs are unambiguously either intransitive or transitive. Transitive verbs are derived from an intransitive stem by suffixation and vice versa. There are also cases where both intransitive and transitive verbs are derived from the same stem (or root) with different suffixes. In (6), the suffix '-e' adds a transitive function to the intransitive stem 'ak-'.

(6)	a.	doa ga	ak-u	
		door NOM	open-PRS	
		Arg1		
		'The door opens'		
	b.	taroo ga	doa o	ak-e-ru
		Taro NOM	door ACC	open-TRZ-PRS
		Arg0	Arg1	
		'Taro opens the door'		

Intransitive and transitive pairs similar to 'ak-u' and 'ak-e-ru' above are registered as different lexical units in morphological analyzers (e.g., MeCab[6]) as well as Japanese dictionaries even though they share the same stem. However, they share the same semantic role set. This is why we need to assign the same semantic role set and frame (that is, frame-based semantic roles) to each verb.

Thus, we define a frame that indicates a shared concept of predicates that belong to the frame, and then the frame unites a consistent semantic role set. Each verb meaning of polysemous verbs is fixed by adding example sentences. Since a polysemous verb has several meanings, each verb meaning is connected to a distinct frame. For example, the verb 'hirak-u', shown in (5), has another meaning as in (7).

(7)	gen'ya o	hirak-u
	westerland ACC	cultivate-PRS
	Arg1	Frame: *development*
	'Someone cultivates the westerland'	

The meaning of 'hirak-u' in (7) is assigned to the *development* frame to which verbs such as 'kaitaku-su-ru' (cultivate) and 'kaihatsu-su-ru' (develop) are also assigned.

2.2. PT: Repository of Semantic Roles and Frames for Japanese Predicates

As a base repository of semantic roles and frames for Japanese predicates, we use PT. PT is composed of hierarchical frames for predicates and each frame indicates a shared meaning of predicates whose sense is designated with semantic-role-annotated example sentences.

In previous work (Takeuchi et al., 2010), PT was constructed for about 11,000 Japanese predicates, with about 23,000 annotated example sentences. The semantic roles in PT have conventional role names (e.g., Agent, Theme, Goal) [7]. Thus, by adding PropBank-style semantic roles to the example sentences in PT, we utilize PT as a repository of frames containing frame based semantic roles that are both numbered and named for Japanese predicates.

[5] The following abbreviations are used in glosses: ABL: ablative, ACC: accusative, CAUS: causative, CVB: converb, DAT: dative, GEN: genitive INS: instrumental, NOM: nominative, PASS: passive, POL: politeness, PRS: present, PST: past, TOP: topic, TRZ:transitivizer.

[6] https://taku910.github.io/mecab/.

[7] Actual names are defined in Japanese (see http://pth.cl.cs.okayama-u.ac.jp/). There are about 70 distinct semantic roles.

Figure 1 shows an example of how the verb 'hirak-u' is registered in PT. As a polysemous verb, 'hirak-u' has *open*, *develop*, and *start* meanings. Each word meaning has an example sentence that is annotated with both numbered semantic roles and named semantic roles. The semantic roles are Arg1 and Theme in all of the example sentences[8] in Figure 1.

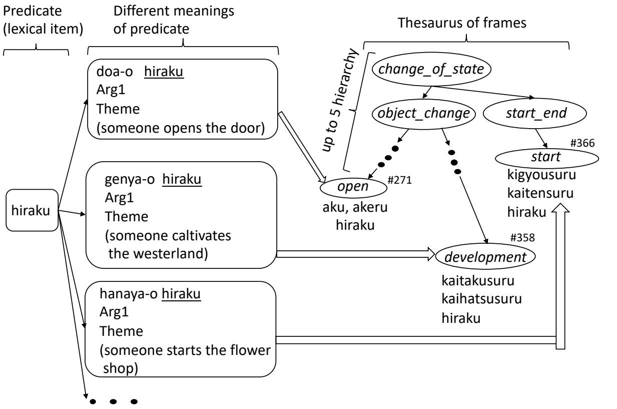


Figure 1: Example of how a polysemous verb is registered in PT, where PT is the repository of semantic roles and frames.

Then each example sentence is linked to a frame in the thesaurus. For example, the meaning of 'hanaya o hirak-u' (someone starts the flower shop) is assigned to *start*[9] whose frame ID is 366. The *start* frame also contains 'kigyoo-su-ru' (start new venture) and 'kaiten-su-ru' (open a shop). The structure of frames is a thesaurus. There is no multiple inheritance in the thesaurus of frames. Such structure is similar to a synset in WordNet.

Frame-based numbered semantic roles are convenient to capture the same type of arguments with different named roles for different verbs in the same frame. For example, both predicate verbs 'oros-u' (get down) and 'ochi-ru' (fall) in (8) and (9) belong to the *moving_from* frame.

(8)	taro ga	hon o	tana kara	oros-u
	Taro NOM	book ACC	shelf ABL	get
	Arg0	Arg1	Arg2	Frame: *moving_from*
	Agent	Theme	Source	

'Taro gets a book down from the shelf'

(9)	chichi ga	yane kara	ochi-ru
	father NOM	roof ABL	fall-PRS
	Arg1	Arg2	Frame: *moving_from*
	Experiencer	Source	

'My father falls from the roof'

In the *moving_from* frame, the mover of the moving event is annotated as Arg0 and Agent. This raises concern for (9), where it is 'chichi' (father) who is moved. For such a case, 'chichi' is annoated with the Experiencer named semantic role because 'chichi' is moved without intending to be moved. According to the annotation guidelines of PropBank (Bonial et al., 2010), Experiencer is typically Arg0 in numbered roles. However, the numbered role Arg1 is usually assigned to the Patient argument, i.e. "the argument which undergoes the change of state or is being affected by the action". Consequently, 'chichi' is annotataed as Arg1 in (9), so as to be consistent in the meaning of something which moves, i.e., 'hon' (book) and 'chichi' (father) in the sentences.

On the other hand, named semantic roles are helpful for understanding the meanings of arguments. The named semantic roles are annotated across frames, and thus the same type of arguments can be extracted, while arguments numbered Arg2 or with a higher number depend on each frame[10].

In (10) and (11), even though the frames are different, 'kangoshi' (nurse) and 'asshoo' (big win) are a complement of the argument with the accusative case marker. For the arguments, the named semantic role is Complement (ACC) that indicates a complement of the argument with accusative case.

(10)	taroo ga	ken o	kangoshi toshite
	Taro NOM	Ken ACC	nurse as
	Arg0	Arg1	Arg2
	Agent	Theme	Complement(ACC)

yato-u
hire-PRS
Frame: *employ*

'Taro hires Ken as a nurse'

(11)	kekka o	asshoo to	happyoo-su-ru
	result ACC	big.win as	announce-do-PRS
	Arg1	Arg2	Frame: *communication*
	Theme	Complement(ACC)	

'Someone announces the result is a big win'

3. Framework of Annotation Task

As mentioned in the previous section, the repository of semantic roles and frames (i.e., PT) is web-accessible, and has search functionality so annotators can look up example sentences in PT. The annotation task is carried out by annotators under the guidance of a supervisor. Figure 2 gives an overview of the annotation task. The procedure of annotation of semantic roles and frames on the NPCMJ is as follows:

1. Target predicates and their arguments are extracted by a program (Treebank Semantics) (Butler, 2019). This works by converting constituency tree annotations into logic based meaning representations from which

[8] Subjects (i.e., Arg0) are often omitted in Japanese, even when the verb is transitive.

[9] To be exact, this is the *change_of_state/start_end/start* frame. For simplicity, we often omit to write the hierarchy of the frame.

[10] From the standpoint of Construction Grammar (Goldberg, 1995), the named semantic roles and the numbered semantic roles can be regarded as 'argument roles' and 'patient roles', respectively. The named semantic roles are annotated according to syntactic expressions of arguments for verbs, while the numbered semantic roles are annotated based on frames.

predicate-argument information can be extracted and remapped back into the tree structures of the NPCMJ for identifying target predicates and their arguments.

2. Annotators select the target sentence, look up a target predicate for an example, find the most appropriate frame for the example, and then assign the frame ID number to the target predicate.

3. Then, annotators assign semantic roles to the arguments according to the examples in PT.

4. If annotators find cases that are missing from PT or annotators are not confident, annotators write a report.

5. The supervisor adds new examples to PT or has discussions with annotators according to their reports[11].

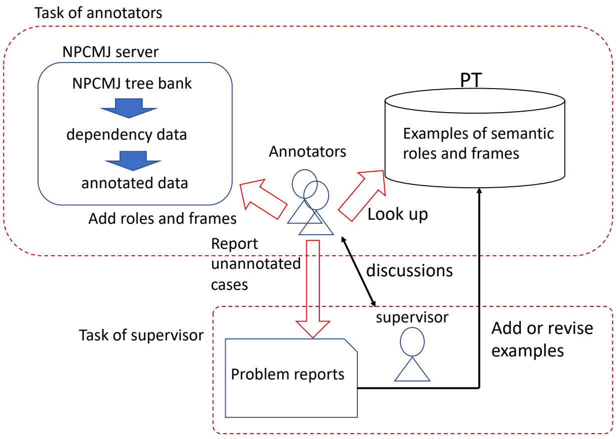
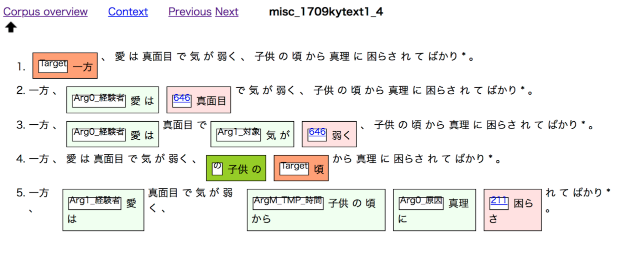


Figure 2: Framework of annotation task.

In step **2**, annotators can see the dependencies of the target sentence on the NPCMJ server (Figure 3). Since the NPCMJ is a web-accessible treebank, annotators can also look at the tree view of the target sentence from the same webpage (Figure 4). As can be seen in Figure 2, annotators input frame IDs and semantic roles directly into the NPCMJ server, then the annotated results can be checked by all the other annotators and the supervisor.


Corpus overview Context Previous Next misc_1709kytext1_4

1. Target 一方 、 愛 は 真面目 で 気 が 弱く 、 子供 の 頃 から 真理 に 困らさ れ て ばかり * 。
2. 一方 、 Arg0_経験者 愛 は 646 真面目 で 気 が 弱く 、 子供 の 頃 から 真理 に 困らさ れ て ばかり * 。
3. 一方 、 Arg0_経験者 愛 は 真面目 で Arg1_対象 気 が 646 弱く 、 子供 の 頃 から 真理 に 困らさ れ て ばかり * 。
4. 一方 、 愛 は 真面目 で 気 が 弱く 、 の 子供 の Target 頃 から 真理 に 困らさ れ て ばかり * 。
5. 一方 、 Arg1_経験者 愛 は 真面目 で 気 が 弱 く 、 ArgM_TMP_時間 子供 の 頃 から Arg0_原因 真理 に 211 困ら さ れ て ばかり * 。

Figure 3: Dependency view on the Web site.

The frames of PT were constructed on the basis of the Lexeed database (Fujita et al., 2006). This database is a sense repository for a rule-based machine translation system. It follows that the frames of PT have a wide coverage for frequent predicates. However, there remain words and word meanings that have not been registered. Adding new words or example sentences to PT is limited to the supervisor only so as to maintain the consistency of the dictionary. A problem with this setup is that the operation of extending PT could become a bottleneck for the annotation.

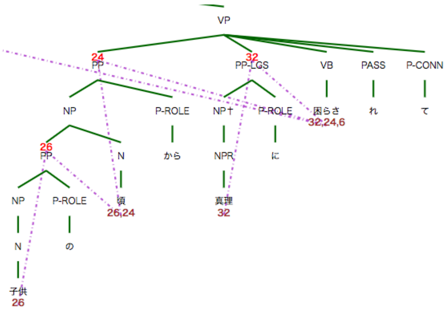


Figure 4: Tree view on the Web site.

4. Examples of Complicated Annotations

With the annotation task, we found syntactic and grammatical variations because, as a corpus, the NPCMJ is composed of texts from very diverse genres, such as white papers, novels, speech dictations, news papers, and so on. Here we show some complicated cases of semantic role annotations.

Causative form
As described in Section 2.1., some transitive verbs have corresponding intransitive verbs. Both verbs can take the causative form. In PT numbered arguments and named roles are assigned on the basis of active voice. Note that causative forms of intransitive verbs exhibit structure similar to transitive verbs.

(12) a. machi ga hatten-shi-ta
city NOM development-do-PST
Arg1 Frame: *development*
Theme
'The city has grown'

b. shichoo ga machi o hatten-sa-se-ta
mayor NOM city ACC development-do-CAUS-PST
Arg0 Arg1 Frame: *development*
Agent Theme
'The Mayor developed the city'

(13) a. taroo ga hon o kai-ta
Taro NOM book ACC write-PST
Arg0 Arg1 Frame: *create*
Agent Theme
'Taro wrote a book'

[11] Currently there are three annotators working on the project. Discussions with the supervisor are used to settle all cases where annotators are not confident.

b.

shuppansha ga	taroo ni	hon o
publisher NOM	Taro DAT	book ACC
ArgA	Arg0	Arg1
Causer	Causee	Theme

kak-ase-ta
write-CAUS-PST
Frame: *create*

'The publisher had Taro write a book'

(14)

funbat-te	kono ugoki o
try.hard-CVB	this movement ACC
ArgM_ADV	Arg1
ADV	Theme

hatten-sa-se-ru	sutamina ga	...
development-do-CAUS-PRS	stamina NOM	
Frame: *development*	ArgA	
	Causer	

'stamina for developing this movement by trying hard is.. (81_ted_talk_5)

Passive form

Japanese has several special grammatical forms in the passive voice. The first one is the adversative passive (Wierzbicka, 1979), that is, the passive form for an intransitive verb.

(15)

ai wa	mari no taido ni	komar-u
Ai TOP	Mari GEN attitude DAT	have.trouble-PRS
Arg1	Arg0	Frame: *suffering*
Experiencer	CAUSE	

'Ai is troubled by Mari's attitude'

This intransitive verb[12] can take an indirect passive after taking a causative form.

(16) a.

mari ga	ai o
Mari NOM	Ai ACC
Arg0	Arg1
Agent	Theme(Person)

komar-ase-ru
have.trouble-CAUS-PRS
Frame: *suffering*

'Mari troubles Ai'

b.

ai wa	mari ni
Ai TOP	Mari DAT
Arg1	Arg0
Experiencer	CAUSE

komar-as-are-ru
have.trouble-CAUS-PASS-PRS
Frame: *suffering*

'Ai is troubled by Mari'

This final causative-passive form is the example in the 5th case of Figure 3.

(17)

ai wa	kodomo no koro kara	mari ni
Ai TOP	child GEN time ABL	Mari DAT
Arg1	ArgM_TMP	Arg0
Experiencer	Time	CAUSE

komar-as-are-te	bakari
have.trouble-CAUS-PASS-CVB	always
Frame: *suffering*	

'Ai is always troubled by Mari since she was a child' (4_misc_1709kytext1)

Light verbs

Japanese often uses the light verb construction. For this construction, there is an argument that works as the predicate, and so this argument that provides the predicate content is withdrawn from being assigned a semantic role. For example, consider (18).

(18)

(pro)	mainichi	suukai	jissaini
	everyday	several.times	actually
Arg0	ArgM_TMP	ArgM_ADV	ArgM_ADV

oshaberi o	shi-mas-u
chatting ACC	do-POL-PRS
ArgM_PRX	Frame: *chat: oshaberi-su-ru*

'(pro) chat everyday several times actually' (47_ted_talk_8)

In (18), the target verb 'shi-ma-su' (do) is a light verb. The content of the action is expressed by the deverbal noun 'oshaberi' (chatting). For this light verb case, the Arg_PRX tag is assigned to the content argument 'oshaberi o'. This follows the treatment proposed in the PropBank guidelines (Bonial et al., 2010). However, the composed meaning of the light verb construction, i.e., the *chat* frame and the verb 'oshaberi-sur-u' (chat-do-PRS) are assigned to the light verb 'shi-ma-su'. Thus the role sets of the *chat* frame are applied to the arguments and adjuncts except for 'oshaberi o' (ArgM_PRX).

5. Current Annotation Results

In this section we discuss the current state of annotation. First, we show the statistics of annotated numbers of sentences, target predicates, types of predicates, and semantic roles. Second, we apply the annotated corpus to a machine learning system to estimate roughly the quality of annotated semantic roles with comparison to previous studies. Currently we are completing the first round of annotation that we plan to review.

5.1. Statistics of Annotated Data

Table 1 shows statistics for what has been annotated.

Annotated sentences	32,044
Annotated predicates	9,878
Annotated arguments	21,454
Type of predicates	2,868

Table 1: Statistics of annotated data

The top 10 most frequently numbered semantic roles are shown in Table 2.

[12] As shown in (15), Experiencer is assigned to Arg1 in cases involving a psychological verb. This follows the treatment of the verb *annoy* in its PropBank frame file.

Arg	A1	A0	A2	M-ADV	M-TMP
#	8357	5413	3733	1209	767
	M-LOC	M	M-CAU	A3	M-PRP
	711	353	204	187	157

Table 2: Statistics of top 10 numbered semantic roles

Table 3 shows the top 10 most frequently named semantic roles.

Role	Th	Agent	Adv	Exp	Temp
#	7470	4028	1363	1212	849
	Loc	Th(ACT)	Goal	Dom	Th(Gen)
	541	486	390	377	315

Table 3: Statistics of top 10 named semantic roles

5.2. Preliminary Experiments of Semantic Role Labeling Using Deep Learning Model

Motivation

Previous work has constructed a semantic role labeling system with neural network models for Japanese (Okamura et al., 2019). The target data is BCCWJ-PT, where data is annotated with the semantic roles defined in PT. While the sentences in BCCWJ-PT are different from those of the NPCMJ, we can roughly estimate the quality of the named semantic roles in the NPCMJ by comparing the accuracy of a semantic role labeling system using the NPCMJ to the case of using BCCWJ-PT.

Semantic role labeling system

The input of the semantic role labeling system is the target predicate and morpheme sequence of its argument (or adjunct), and then the output is a semantic role label of the argument. All of the morphemes are converted to d dimension vectors with nwjc2vec, which gives Japanese word vectors[13]. Let $\boldsymbol{X}$ be a vector sequence of an input morpheme sequence, that is, $\boldsymbol{X} = \boldsymbol{x}^1, \ldots, \boldsymbol{x}^t, \ldots, \boldsymbol{x}^T$ ($\boldsymbol{x}^t \in \mathbb{R}^d$). Let S be output of a semantic role label for input $\boldsymbol{X}$, the estimated semantic role label is defined by Formula (1).

$$\hat{S} = \underset{j \in Sem}{argmax}\, p(S_j|\boldsymbol{X}) \tag{1}$$

Where, Sem is a set of semantic role labels, and $S_j (j = 1, \ldots, Sem)$ is the jth semantic role label. Let y_j be an output of the jth unit at the final output layer of the neural network model. Since the softmax function is used as a non-linear function at the output layer, y_j can be a probability.

$$y_j = p(S_j|\boldsymbol{X}). \tag{2}$$

Then $\hat{S}$ can be estimated by using a neural network.

As a neural network model, we apply bi-directional GRU with the max-pooling model (hereafter referred to as the bi-GRU model) to the semantic role labeling because the bi-GRU model gave the best performance in the previous study of Okamura et al. (2019). Figure 5 shows the architecture of the bi-GRU model. An input vector sequence $\boldsymbol{X}$ is applied to the input of bi-directional GRU, and then the max-pooling is applied to outputs of the GRU with time sequence direction. The final output $\boldsymbol{y} = [y_1, \ldots, y_j, \ldots, y_{Sem}]$ is obtained after applying a fully-connected layer to the results of max-pooling.

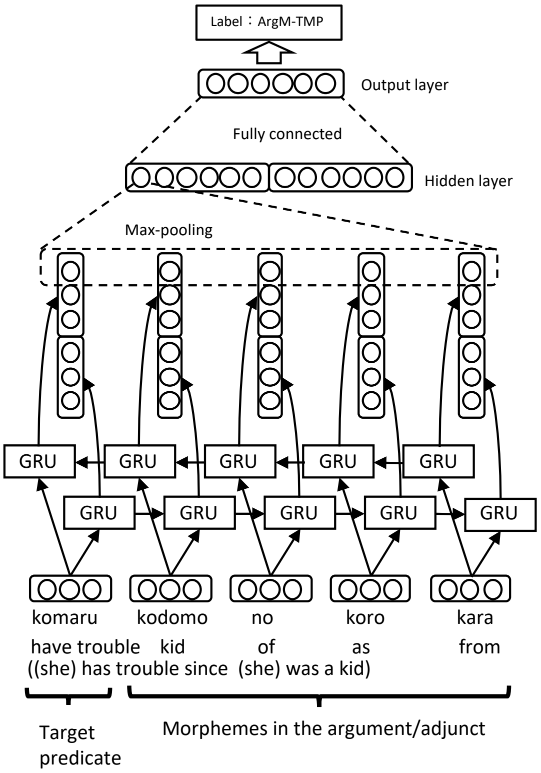


Figure 5: Bi-directional GRU with max-pooling model

Experimental setup

The number of hidden units of GRU is 256. The settings of the optimizer are set the same as in Okamura et al. (2019). The annotated data is divided into 65%, 5%, and 30% for training, development, and test data, respectively. The performance is evaluated by the accuracy of the test data.

$$\text{Accuracy} = \frac{\text{\# Estimated semantic roles are correct}}{\text{\# All instances}} \tag{3}$$

Experimental results

Table 4 shows the total accuracies of numbered and named semantic role labels. The accuracy of the third column shows the results for BCCWJ-PT (Okamura et al., 2019). According to the accuracies of named semantic roles in Table 4, the accuracy of the model using NPCMJ is near to that of BCCWJ-PT. BCCWJ-PT was annotated with two annotators for each semantic role as well as being checked

	NPCMJ	BCCWJ-PT
Numbered semantic roles	0.716	N/A
Named semantic roles	0.667	0.702

Table 4: Total accuracy of semantic roles

[13] http://nwjc-data.ninjal.ac.jp/

by a third annotator, and so the quality of annotated semantic roles is expected to be high. The above results indicate that for the current annotated semantic roles of the NPCMJ the consistency of the annotated tags is promising for a resource that is under development.

Comparing numbered and named semantic roles in their accuracy, numbered semantic roles are higher than named semantic roles. The result can be considered to show that the numbered semantic roles are annotated more consistently than the named semantic roles. Table 5 and Table 6 show accuracies for both numbered roles and semantic roles of the top 10 most frequent.

Arg	A1	A0	A2	M-ADV	M-TMP
Acc	0.818	0.822	0.591	0.647	0.661
	M-LOC	M	M-CAU	A3	M-PRP
	0.479	0.095	0.556	0.078	0.167

Table 5: Accuracies of the top 10 most frequent numbered semantic roles

Role	Th	Agent	Adv	Exp	Temp
Acc	0.826	0.855	0.672	0.446	0.745
	Loc	Th(ACT)	Goal	Dom	Th(Gen)
	0.506	0.38	0.574	0.226	0.207

Table 6: Accuracies of the top 10 most frequent named semantic roles

Comparing the accuracy of numbered semantic roles to the results for the shared task in English semantic role labeling tasks, we have to improve our semantic role labeling models.

6. Discussions

One of the difficulties of annotating semantic roles is to discriminate between arguments and adjuncts, while keeping frame consistency in PT. Arguments are the essential factors for the defined frame, while adjuncts are not core elements for a frame. In principle, adjuncts can be attached to any frame.

As described in Section 2.1. we added Arg0, Arg1, etc. tags to the previously defined named semantic roles in PT for the frame repository. However, because of the lack of variation in the example sentences, we have found arguments to be missing. This applies especially to arguments that are inserted as parts of constructions.

Consider (19), and the need for a *create* frame that contains the meaning of the verb 'kak-u' (write). For (19), we need to define two essential semantic roles, namely: Arg0 (Agent, writer) and Arg1 (Theme, written thing).

(19)	ji o	gayoushi ni	kak-u
	character ACC	drawing.paper DAT	write-PRS
	Arg1	ArgM_LOC	Frame: *create*
	Theme(Gen)	Location	

'Someone writes characters in the drawing paper' (dict_pth_c_825)

When the argument Arg1 is 'ji o' (character ACC), there is no expectation for other essential arguments. When the argument Arg1 is'tegami' (letter), then the recipient of the letter 'imooto ni' (to sister) can also be present, as in (20).

(20)	tegami o	imooto ni	kak-u
	letter ACC	sister DAT	write-PRS
	Arg1	Arg2?	Frame: *create*
	Theme	Theme (Person)	

'Someone writes a letter to someone's sister'

The recipient 'imooto-ni' (sister DAT) must be part of a construction (Goldberg, 1995). Thus, the recipient can be considered as an argument because the recipient appears depending on the *create* frame. We can also confirm the semantic role of the above 'recipient' case by looking at the corresponding examples in English PropBank and FrameNet.

In the frameset *write.01* of PropBank, the corresponding roles are defined as A2 (benefactive), that is an essential argument. In FrameNet, the frames are more detailed than our frames in PT. The meaning of the above case, i.e., 'write' in English, is assigned to the *Contacting* frame. In the *Contacting* frame, the recipient role is defined as the *Addressee* role that indicates a core role (i.e., an essential argument). Thus both English language resources offer an analysis that is the same as our analysis for the Japanese data.

Next, consider (21)[14], which is another example that is not registered in PT.

(21)	kyuujitai de wa	shoomyoo to
	old.kanji.form INS TOP	Shomyo as
	Domain	Complement(TOP)

kak-u
write-PRS
Frame: *create*

'(They) write it as "Shomyo" in the old kanji style' (5_wikipedia_KYOTO_11)

We are currently investigating whether the two phrases ('kyuujitai de' and 'shoomyoo to') are arguments and/or adjuncts for the *create* frame of PT. According to FrameNet, the verb 'write' can belong to the *Statement*, *Text creation*, *Contacting* and *Spelling and pronouncing* frames. The *create* in PT can partially correspond to the *Text creation* frame. However, it is the *Spelling and pronouncing* frame that seems to correspond to (21). In the *Spelling and pronouncing* frame, 'kyuujitai de' might correspond to *Manner* and 'shoomyoo to' to *Formal_realization*; and both roles are defined as core roles.

This is suggestive evidence that FrameNet has analyzed examples that can be expected to cover Japanese examples too, even though Japanese has different syntactic and grammatical characteristics when compared to English.

Thus we think it will be helpful to refer to existing language resources, especially PropBank and FrameNet, in revising framesets in PT. Such comparisons are only possible

[14] The 'shoomyoo' indicates chanting of Buddhist hymns. The sentence describes the character shape of 'Shomyo'.

because there are web-accessible frame data sets, notably, PropBank and FrameNet. Thus, we believe that providing the NPCMJ and the frame repository PT as web-accessible resources is essential for being able to construct reliable semantic role labeled language resources for Japanese and beyond.

7. Conclusion

This paper has described ongoing research of constructing an annotated corpus of semantic roles and frames as an addition to the NPCMJ, a Japanese web-accessible parsed corpus. The annotation task is coupled with the expansion of PT, that is, the repository of semantic roles and frames. We annotate two types of semantic roles, i.e., numbered and named semantic roles for the arguments. Both the numbered semantic roles and the named semantic roles are defined consistently with respect to frames. Then numbered semantic roles are expected to be used in NLP, and the named roles are to be used as tags for searching example sentences by language learners and linguists. In the annotated texts, we have found various kinds of syntactical and grammatical variations: e.g., adversative passives, alternations of cases, and collocations. We also applied part of the annotated corpus into a semantic role annotation model based on neural networks to evaluate how the annotated corpus can contribute to the statistical learning model approach. The results show that the accuracies of semantic role annotation systems are almost the same as the current quality of named semantic roles, for results near to the achievements of previous work. Thus, we can estimate that the quality of the annotated semantic roles are highly promising. What is currently annotated is the first part of a planned annotation cycle, and so all is due for review as we also continue to annotate new texts from the NPCMJ. The NPCMJ is planned to increase in size to 60,000 sentences.

8. Acknowledgements

A part of the research reported in this paper is supported by the NINJAL project "Development of and Research with a parsed corpus of Japanese" through the NINJAL collaborative research project "Annotating Semantic Role Labels and Frames to NPCMJ", and by JSPS KAKENHI (Grant Number JP 15H03210 and 19K00552).

9. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 86–90.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Bonial, C., Babko-Malaya, O., Choi, J. D., Hwang, J., and Palmer, M., (2010). *PropBank Annotation Guidelines Version 3.0.* (`http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf` accessed 2019/9/6).

Butler, A. (2019). Treebank Semantics. Technical report, Hirosaki University. (`http://www.compling.jp/ajb129/tsdoc.html` accessed 2019/11/30).

EDR, (1995). *EDR: Electric Dictionary the Second Edition.* Japan Electronic Dictionary Research Institute, Ltd.

Fillmore, C. J., (1968). *The Case for Case*, pages 1–89. New York: Holt, Rinehart, and Winston.

Fujita, S., Tanaka, T., Bond, F., and Nakaiwa, H. (2006). An implemented description of japanese: The lexeed dictionary and the hinoki treebank. In *Proceedings of the COLING/ACL06 Interactive Presentation Sessions*, pages 65–68.

Goldberg, A. E. (1995). *Constructions.* The University of Chicago Press.

Hashida, K. (2005). GDA: Japanese Annotation Manual Version 0.74. (`http://i-content.org/gda/tagman.html` accessed 2019/11/30).

Horn, S. W., Butler, A., Nagasaki, I., and Yoshimoto, K. (2018). Derived mappings for FrameNet construction from a parsed corpus of Japanese. In *LREC 2018 Proceedings, International FrameNet Workshop, 11th edition of the Language Resources and Evaluation Conference*, pages 28–32, Miyazaki, Japan.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese Text Corpus with a Predicate-Argument and Coreference Relations. In *Proceedings of the 1st Linguistic Annotation Workshop*, pages 132–139.

Kawahara, D., Kurohashi, S., and Hasida, K. (2002). Construction of a japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 2008–2013.

Loper, E., ting Yi, S., and Palmer, M. (2007). Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, pages 118–128.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.

NINJAL. (2016). NINJAL Parsed Corpus of Modern Japanese. (Version 1.0). Technical report, National Institute for Japanese Language and Linguistics. (`http://npcmj.ninjal.ac.jp/interfaces/` accessed 2019/11/30).

Ohara, K. H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., and Ishizaki, S. (2006). Frame-based contrastive lexical semantics and japanese framenet: The case of risk and kakeru. In *Proceeding of the Fourth International Conference on Construction Grammar.* http://jfn.st.hc.keio.ac.jp/ja/publications.html.

Okamura, T., Takeuchi, K., and Ishihara, Y. (2019). Using Neural Networks to Construct a Japanese Semantic Role Labeling Model. *Journal of Information Processing*, 60:2063–2074. (in Japanese).

Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling*. Morgan & Claypool Publishers.

Takeuchi, K., Inui, K., Takeuchi, N., and Fujita, A. (2010). A Thesaurus of Predicate-Argument Structure for Japanese Verbs to Deal with Granularity of Verb Meanings. In *The 8th Workshop on Asian Language Resources*, pages 1–8.

Takeuchi, K., Ueno, M., and Takeuchi, N. (2015). Annotating Semantic Role Information to Japanese Balanced Corpus. In *Proceedings of MAPLEX 2015*.

Wierzbicka, A. (1979). Are Grammatical Categories Vague or Polysemous? (The Japanese ‘ Adversative ’ Passive in a Typological Context). *Paper in Linguistics*, 12(1-2):111–162.

Zhang, S., Ma, X., Duh, K., and Durme, B. V. (2019). Amr parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.